DOCUMENT RESUME

ED 463 320 TM 033 752

AUTHOR Kennedy, Charlotte A.

TITLE The Sampling Distribution and the Central Limit Theorem: What They Are and Why They're Important.

PUB DATE 2002-02-15

NOTE 20p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (Austin, TX, February 14-16, 2002).

PUB TYPE Information Analyses (070) -- Reports - Evaluative (142)

EDRS PRICE MF01/PC01 Plus Postage.

DESCRIPTORS Educational Research; Hypothesis Testing; Research Methodology; *Sampling; Statistical Distributions; *Statistical Significance

IDENTIFIERS *Central Limit Theorem

ABSTRACT

The use of and emphasis on statistical significance testing has pervaded educational and behavioral research for many decades in spite of criticism by prominent researchers in this field. Much of the controversy is caused by lack of understanding or misinterpretations. This paper reviews criticisms of statistical significance testing and discusses concepts related to the sampling distribution and the central limit theorem and the role these concepts play in statistical significance testing. Statistical significance testing should not be used as the sole basis for analyzing hypotheses of scientific inquiry, but it can be an effective tool when used with good judgment. (Contains 22 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made from the original document.

Running Head: THE SAMPLING DISTRIBUTION

ED 463 320

The Sampling Distribution and the Central Limit Theorem: What They are and Why They're Important

Charlotte A. Kennedy

Texas A&M University

C. Kennedy

U.S. DEPARTMENT OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM033752

Paper presented at the annual meeting of the Southwest Educational Research Association, Austin, TX, February 15, 2002.

BEST COPY AVAILABLE

Abstract

The use of and emphasis on statistical significance testing has pervaded educational and behavioral research throughout many decades despite staunch criticisms by prominent researchers in this field. The lack of understanding and misinterpretations of statistical significance cause much of the controversy. Therefore, this paper reviews numerous criticisms with statistical significance testing as well as discusses concepts related to the sampling distribution and the central limit theorem and their role with statistical significance testing.

The Sampling Distribution and the Central Limit Theorem: What They are and Why They're Important

According to Huberty (1993), statistical significance testing dates back nearly 300 years to studies conducted by John Arbuthnot in 1710. Although its use is prevalent throughout the behavioral social sciences, the efficiency and advantages of statistical significance testing is questionable (Carver, 1978; Cohen, 1994; Kirk, 1996; Nickerson, 2000; Thompson, 1993, 1999a, 1999b). Oftentimes researchers lack the basic understanding of the principles of statistical significance testing, thereby causing misinterpretations in their results (Carver, 1978; Kirk, 1996; Thompson, 1994). As Thompson (1996) stated, "many people who use statistical test might not place such a premium on the test if these individuals understood what the tests really do, and what the tests do not do" (p. 26). Therefore, this article will begin by then discussing the fundamentals of statistical significance testing and the sampling distribution and proceed by reviewing various criticisms and suggestions when using statistical significance testing.

What Statistical Significance Tests Do

Stated simply, to obtain "statistical significance," the $p_{calculated}$ must be less than the $p_{critical}$ ($p_{calculated} < \alpha$). However, the principles underlying statistical significance testing are more complex. Although $p_{critical}$ is a subjective choice made by the

researcher (usually set at .01 or .05) and ultimately indicates how "scared" the researcher is of making a Type I error (Thompson, 1994, 1999b), the concept of $p_{calculated}$ is more complex. Thompson (1996) defined $p_{calculated}$ as the "probability (0-1.0) of the sample statistics, given the sample size, and assuming the sample was derived from a population in which the null hypothesis (H_0) is exactly true" (p. 27). From the definition of $p_{calculated}$ provided by Thompson (1996), several points regarding statistical significance need to be highlighted.

What Statistical Significance Tests Do Not Do

For one, $p_{calculated}$ is the probability of the sample statistics assuming (not testing) the population parameters (Thompson, 1996). Therefore, although researchers wish to generalize their results to the population of study, when statistical significance testing is performed, the direction of inference is from the population to the sample, not from the sample to the population (Thompson, 1998). In addition to sample statistics and population parameters, sample size plays a key role in whether or not statistically significant results will be found (Carver, 1978; Thompson, 1996). With a large enough sample size, the null hypothesis will always be rejected and statistical significance will be obtained (Carver, 1978; Cohen, 1990, 1994; Kirk, 1996; Nickerson, 2000; Thompson, 1996, 1998). As Thompson (1998) asserted, if a researcher is not able

to reject the null hypothesis and thus find statistical significance, then that researcher was "too lazy to drag in enough participants" (p. 799). Ultimately, the null hypothesis will always be rejected given a large enough sample size (Cohen, 1990). Therefore, the assumption underlying $p_{calculated}$ is inherently flawed in its "assuming the sample was derived from a population in which the null hypothesis (H_0) is exactly true" because the null hypothesis arguably will never be exactly true in the population. As Cohen (1990) pointed out, "So if the null hypothesis is always false, what's the big deal about rejecting it?" (p. 1308).

Several limitations and "false beliefs" (Nickerson, 2000) regarding the interpretations from statistical significance testing exist. For example, Schmidt (1996) cautioned researchers and readers that the binary decision of whether to reject or to not reject the null hypothesis promotes the erroneous idea that if the null is rejected then it must be true and vice versa. As Nickerson (2000) noted, failing to reject the null is not the same as demonstrating it to be true because the null will be rejected with a large enough sample size. Therefore, as previously stated, the null hypothesis is always false (Carver, 1978; Cohen, 1990, 1994; Kirk, 1996; Nickerson, 2000; Thompson, 1998) and "if the null hypothesis is never true, then evidence that it should be rejected in any particular

instance is neither surprising nor useful" (Nickerson, 2000, p. 266) and "in fact, [if you do not reject the null hypothesis,] all you could conclude is that you *couldn't* conclude that the null was false." (Cohen, 1990, p. 1308). No further interpretations can be made. The results from the statistical significance testing should not be the only means to determine if the study is worthwhile.

Even if "statistical" significance is found (the null hypothesis was rejected), the implications of the results do not necessarily warrant "practical" significance, which can often be revealed by the effect size (Kirk, 1996; Rosnow & Rosenthal, 1989; Thompson, 1996; Schmidt, 1996), or "clinical" significance (Thompson, 2002). To note, however, Nickerson (2000) also warns that a "large effect is not a guarantee of importance any more than a small p-value" (p. 257). In other words, a small $p_{calculated}$ value or a large effect size does not necessarily indicate that the results are important to "real-world" application (Nickerson, 2000). Furthermore, to avoid confusion in this interpretation of the results, the phrase "statistically significant" should be employed instead of simply "significant" (Carver, 1978; Nickerson, 2000; Thompson, 1994, 1996). "Significant" implies "important" and, again, the statistically significant results may not be necessarily important in reality.

To determine the practical significance of the obtained results, all factors, including personal values, must be examined.

As previously mentioned, because the direction of inference is from the population to the sample (Cohen, 1994; Thompson, 1998), statistically significant results should not be interpreted to suggest that the results are replicable (Carver, 1978; Nickerson, 2000; Schmidt, 1996; Thompson, 1996). As Carver (1978) stresses, "Statistical significance simply means statistical rareness" (p. 383). The only interpretation of the results that can be presented when statistical significance is reached is that the results are unlikely given the sample size and assuming that the null hypothesis is exactly true in the population (Carver, 1978; Thompson, 1994).

Because statistical significance testing does not signify that the results are replicable, alternative methods must be employed, which may consist of external or internal techniques. External methods of examining replicability are the actual, physical replication of the study with a different sample (Thompson, 1996). For the sake of time and convenience, however, most researchers prefer internal analyses to confirm result replication and examples of these analyses include cross-validation, jackknife, and the bootstrap (Thompson, 1996, 1993). These internal analyses involve manipulation via different groupings but are limited because "all yield somewhat inflated

estimates of replicability" (Thompson, 1993, p. 368). But, as Thompson (1993) further indicated, it is best to have some conception of the replicability of the results than none at all.

Because of the many criticisms regarding statistical significance, the American Psychological Association (APA) created a Task Force on Statistical Inference to review these criticisms as well as other issues regarding what to report in publications (Wilkinson & APA Task Force, 1999). Although this task force has considered banning statistical significance reports in APA journals, Thompson (1993) notes that statistical significance testing should not be completely prohibited but should be recognized as being of "limited value and should not be over interpreted and that these tests can be usefully augmented by analyses that bear more directly on the cumulation of knowledge" (p. 378).

Now the fifth edition of the APA (2001) Publication Manual has been released. The new edition goes considerably beyond the previous edition's "encouragement" (p. 18) to report effect sizes. The new manual emphasizes:

> For the reader to fully understand the importance of your findings, it is almost always necessary to include some index of effect size or strength of relationship in your Results section. You can estimate the magnitude of effect or the strength

> of the relationship with a number of common effect size estimates... The general principle to be followed... is to provide the reader not only with information about statistical significance but also with enough information to assess the magnitude of the observed effect or relationship. (pp. 25-26, emphasis added)

Both before and after the release of the new manual, journal editors began adopting *requirements* that authors report effect sizes as indices of "practical" significance. The following 17 journals now require the reporting of effect sizes:

Career Development Quarterly

Contemporary Educational Psychology

Educational and Psychological Measurement

Exceptional Children

Journal of Agricultural Education

Journal of Applied Psychology

Journal of Community Psychology

Journal of Consulting & Clinical Psychology

Journal of Counseling and Development

Journal of Early Intervention

Journal of Educational and Psychological Consultation

Journal of Experimental Education

Journal of Learning Disabilities

Language Learning

Measurement and Evaluation in Counseling and Development

The Professional Educator

Research in the Schools.

This list includes the flagship journals of the American Counseling Association (distributed to all 55,500+ members) and the Council for Exceptional Children (distributed to all 55,000+ members). Therefore, statistical significance testing should not be banned but used as a supplemental source to obtain a more comprehensive interpretation of statistical analyses. In another article, Thompson (1996) noted:

> We must understand the bad implicit logic of person who misuse statistical tests if we are to have any hope of persuading them to alter their practices-it will not be sufficient merely to tell researchers not to use statistical tests, or to use them more judiciously. (p. 26)

Therefore, although several criticisms and limitations of statistical significance testing have been presented above as well as a basic insight to concepts related to $p_{calculated}$ and $p_{critical}$, the essential foundation related to statistical significance and the sampling distribution has not yet been discussed.

Sampling Distribution

Ultimately, to understand statistical significance testing, one must also understand how $p_{calculated}$ is derived. Therefore, the concepts related to the sampling or underlying distribution are critical to understand because where the statistic is located on the curve tells us the probability of $p_{calculated}$ and whether the $p_{calculated}$ is less than or greater than the $p_{critical}$. The sampling distribution is derived from taking all possible samples and computing all possible statistics and graphing their frequency distribution on a histogram. As with a normal distribution, the area of a sampling distribution is equal to 1.00 (Hinkle, Wiersma, & Jurs, 1998).

Although the population and sample both contain individual scores, the sampling distribution contains statistics. The sampling distribution differs depending upon the different scores or sample sizes extracted as well as with different statistics (such as the mean, median, kurtosis, etc.) employed. The only instance when the sampling distribution has scores is when the sample size is equal to one; then, the mean of a given sample (when n=1) is equal to the score of that same sample (Lewis, 2000). The sampling distribution is then equal to the population score distribution.

Furthermore, the standard deviation of the sampling distribution is known as the "standard error of the statistic"

in the sampling distribution, as opposed to simply the "standard deviation of the sampling distribution" (Hinkle et al., 1998). Like the standard deviation of a sample or population, the standard error reveals the "spread-outism" of the sample statistics in the sampling distribution (Lewis, 2000). In the case where n=1, the standard deviation of the population and the standard error of the sampling distribution are equal. However, when the sample size is infinitely large, then the standard error is closest to zero (for graphical representation, see Hinkle et al., 1998, p. 177).

To establish whether statistical significance has been reached, Hinkle et al. (1998) highlight these steps: (1) State the hypothesis, (2) Set the criterion for rejecting the null (the $p_{critical}$/alpha level), (3) Compute the test statistic (which is similar to computing the $p_{calulated}$), and (4) Decide whether to reject the null hypothesis (p. 200). Hinkle et al. (1998) define the test statistic as a "standard score indicating the difference between the observed sample mean and the hypothesized value of the population mean" (p. 199). Although these authors refer to the mean, the test statistic is not limited to the mean and any statistic can be used (Hinkle, 1998). Agresti and Finlay (1986) note, "knowledge of the sampling distribution of the test statistic allows us to calculate the probability that

specific values of the statistic (e.g., values such as the one actually observed) would occur" (p. 124).

After the test statistic is computed, the researcher decides whether to reject the null hypothesis based on whether the test statistic is location in the region of rejection (Arney, 1990; Hinkle et al, 1998). This information ultimately informs the researcher how unlikely their test statistic is if the null hypothesis were true (Agresti & Finlay, 1986). The direction of the null hypothesis and the $p_{critical}$ determines the region of rejection. For example, if the $p_{critical}$ is set at the .05 level and the test is non-directional (or two-tailed), then .025 or 2.5% of both sides or tails of the distribution will be the region of rejection. However, if the $p_{critical}$ remains stable at the .05 level and the test is directional (or one-tailed), then 5% of one side will be the region of rejection. The side or tail for this directional type of hypothesis that will be the region of rejection depends on the target of the hypothesis. Regardless of the direction, if the test statistic is located in the region of rejection once computed, then the null hypothesis will be rejected and "statistical significance" will be obtained. However, if the test statistic does not fall in the region of rejection, then the null hypothesis will not be rejected and statistical significance will not be reached (Arney, 1990; Hinkle et al., 1998).

Central Limit Theorem

To account for biasness in statistical estimates (because in inferential statistics, estimates are often used to approximate population parameters), mathematical theorems are developed to describe the shapes, central tendencies, and "spread-outism" of sampling distributions (Hinkle et al., 1998). An example of such a theorem that describes the shape is the central limit theorem, which states that as the sample size increases, the sampling distribution becomes more normal even when the population is not normal or is skewed (Hinkle et al., 1998; Mittag, 1992). In those cases in which the shape of the distribution is unknown, Thompson (1993) also mentions that the bootstrap method not only provides the researcher with an estimate regarding result replicability but can also be employed to reveal whether the sampling distribution is not normal. In regards to sample size, Carver (1978) also notes "the average sampling error [or "flukiness"] becomes smaller as the size of the sample becomes larger and it also becomes smaller as the variation of the numbers in the population gets smaller" (p. 38).

In addition to the central limit theorem, unbiased estimators approximate the central tendencies and "spread-outism" of the sampling distribution. If the statistic of the sampling distribution is equal to that of the population, then

the statistic is considered to be an "unbiased estimator" (Hinkle et al., 1998). Hinkle et al. (1998) therefore proclaim that over time and with several replications, then the unbiased estimator will eventually equal the population parameter. Again, replications of the results are necessary to ensure accuracy of the results.

Summary

To conclude, statistical significance testing should not be used as the sole basis for analyzing hypotheses of scientific query. However, although it has its limitations and numerous criticisms, statistical significance testing should not be banned from behavioral science publications. As stated by Nickerson (2000), statistical significance testing can be an effective tool when used with good judgment. Yet, to gain the most comprehensive picture of the data, as much information as possible should be presented in all publications, whether via statistical significance testing, effect sizes, confidence intervals, or internal replicability results. Ultimately, researchers need to be aware of the analyses they are running as well as know how to accurately interpret their results. Until then, many studies of behavioral and social sciences will continue to be corrupted with erroneously applied and misinterpreted statistical tests.

References

Agresti, A., & Finlay, B. (1987). *Statistical methods for the social sciences* (2nd ed.). San Francisco, CA: Dellen Publishing Company.

American Psychological Association (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: Author.

Arney, W.R. (1990). *Understanding statistics in the social sciences.* New York: W.H. Freeman and Company.

Carver, R.P. (1978). The case against statistical significance Testing. *Harvard Educational Review, 48,* 378-399.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45,* 1304-1312.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49,* 997-1003.

Kirk, R.E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56*(5), 746-759.

Hinkle, D.E., Wiersma, W., & Jurs, S.G. (1998). *Applied statistics for the behavioral sciences* (4th ed.). Boston, MA: Houghton Mifflin Company.

Huberty C.J. (1993). Historical origins of statistical testing

Practices: The treatment of Fisher versus Neyman-Pearson views in textbooks. *Journal of Experimental Educations, 61*, 317-333.

Lewis, C.P. (1999, January). *Understanding the sampling distribution and the central limit theorem.* Paper presented at the annual meeting of the Southwest Educational Research Association, San Antonio, TX. (ERIC Document Reproduction Service No. ED426100)

Mittag, K. (1992, April). *Using computers to teach the concepts of the central limit theorem.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco. (ERIC Document Reproduction Service No. ED349947)

Nickerson, R.S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5*, 241-301.

Rosnow, R.L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Pscyhologist, 44*, 1276-1284.

Schmidt, F. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for the training of researchers. *Psychological Methods, 1*(2), 115-129.

Thompson, B. (1993). The use of statistical significance tests

in research: Bootstrap and other alternatives. *Journal of Experimental Education, 61,* 361-377.

Thompson, B. (1994). *The concept of statistical significance testing.* Washington, DC: The Catholic University of America, Department of Education. (ERIC Document Reproduction Service No. ED3666554)

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher, 25*(2), 26-30.

Thompson, B. (1998). In praise of brilliance: Where that praise really belongs. *American Psychologist, 53,* 799-800.

Thompson, B. (1999a). If statistical significance tests are broken/misused, what practices should supplement or replace them?. *Theory & Psychology, 9*(2), 165-181.

Thompson, B. (1999b). Statistical significance tests, effect size reporting, and the vain pursuit of pseudo-objectivity. *Theory & Psychology, 9*(2), 191-196.

Thompson, B. (2002). "Statistical," "practical," and "clinical": How many kinds of significance do counselors need to consider? *Journal of Counseling and Development, 80,* 64-71.

Wilkinson, L., & APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals:

Guidelines and explanations. *American Psychologist, 54,* 594-604.